涉密论文 □　公开论文 ☑

浙 江 大 學

本科生毕业论文

题目　知识驱动的物联网嵌入式固件自动化根因分析方法研究

姓名与学号　张乔 3200102817

指导教师　纪守领

年级与专业　2020级 计算机科学与技术

所在学院　计算机科学与技术学院

递交日期　递交日期

浙江大学本科生毕业论文（设计）承诺书

1. 本人郑重地承诺所呈交的毕业论文（设计），是在指导教师的指导下严格按照学校和学院有关规定完成的。

2. 本人在毕业论文（设计）中除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得**浙江大学**或其他教育机构的学位或证书而使用过的材料。

3. 与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

4. 本人承诺在毕业论文（设计）工作过程中没有伪造数据等行为。

5. 若在本毕业论文（设计）中有侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

6. 本人完全了解**浙江大学**有权保留并向有关部门或机构送交本论文（设计）的复印件和磁盘，允许本论文（设计）被查阅和借阅。本人授权**浙江大学**可以将本论文（设计）的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或扫描等复制手段保存、汇编本论文（设计）。

作者签名：　　导师签名：

签字日期：　　年　月　日　　签字日期　　年　月　日

致 谢

摘 要

Abstract

目录

第一部分　毕业论文

第二部分　毕业论文开题报告

第一部分

毕业论文

1 绪论

1.1 背景

1.1.1 节标题

2 Overleaf 使用注意事项

如果你在 Overleaf 上编译本模板，请注意如下事项[zjuthesis]：

- 删除根目录的 ".latexmkrc" 文件，否则编译失败且不报任何错误
- 字体有版权所以本模板不能附带字体，请务必手动上传字体文件，并在各个专业模板下手动指定字体。具体方法参照 GitHub 主页的说明。
- 当前（2019 年 9 月 2 日）的 Overleaf 使用 TexLive 2017 进行编译，但一些伪粗体复制乱码的问题需要 TexLive 2019 版本来解决。所以各位同学可以在 Overleaf 上编写论文，但务必使用本地的 TexLive 2019 来进行最终编译，以免产生查重相关问题。具体说明参照 GitHub 主页。

2.1 节标题

2.1.1 小节标题

我们可以用 includegraphics 来插入现有的 jpg 等格式的图片，如图 2.1。

图 2.1 浙江大学 LOGO

如表 2.1所示，这是一张自动调节列宽的表格。

如式 2-1，这是一个公式

表 2.1 自动调节列宽的表格

第一列	第二列
xxx	xxx
xxx	xxx
xxx	xxx

$$A = \overbrace{(a+b+c)+\underbrace{i(d+e+f)}_{\text{虚数}}}^{\text{复数}} \tag{2-1}$$

如代码 2.1所示，这是一段代码。计算机学院的代码样式可能与其他专业不同，如有需要，可以从计算机学院专业模板中复制相关的代码样式设定。

代码 2.1: simple.c

```
#include <stdio.h>

int main(int argc, char *argv[])
{
    printf("Hello, zjuthesis\n");
    return 0;
}
```

2.2 关于字体

英文字体通常提供了粗体和斜体的组合，中文字体通常没有粗体或斜体，本模板使用了 ‘AutoFakeBold’ 来实现中文伪粗体，但不提供中文斜体，如表 2.2所示。

同一页上的章标题

表 2.2 一些字体示例

字体	常规	粗体	斜体	粗斜体
Times New Roman	Regular	**Bold**	*Italic*	***BoldItalic***
仿宋	常规	**粗体**	斜体	**粗斜体**
宋体	常规	**粗体**	斜体	**粗斜体**
黑体	**常规**	**粗体**	**斜体**	**粗斜体**
楷体	常规	**粗体**	斜体	**粗斜体**

附录

A　一个附录

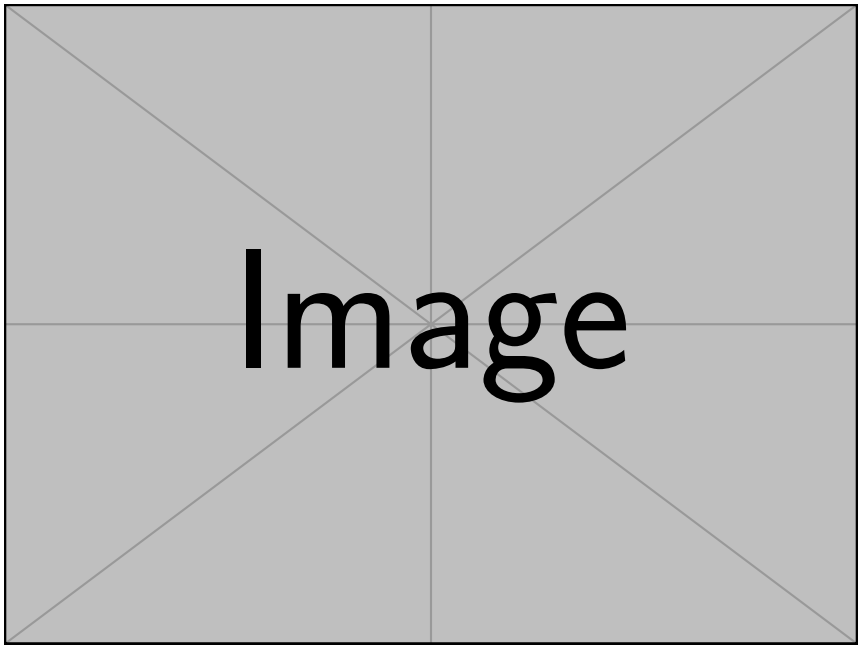

图 A.1 附录中的图片

$$E = mc^2 \tag{A-1}$$

B　另一个附录

$$x^n + y^n = z^n \tag{B-1}$$

作者简历

本科生毕业论文（设计）任务书

一、题目：

二、指导教师对毕业论文（设计）的进度安排及任务要求：

起讫日期 20　年　月　日　至　20　年　月　日

指导教师（签名）_____ 职称 _____

三、系或研究所审核意见：

负责人（签名）_____

年　月　日

本科生毕业论文（设计）考核

一、指导教师对毕业论文（设计）的评语：

指导教师（签名）_____

年　月　日

二、答辩小组对毕业论文（设计）的答辩评语及总评成绩：

成绩比例	文献综述（10%）	开题报告（15%）	外文翻译（5%）	毕业论文质量及答辩（70%）	总评成绩
分值					

负责人（签名）_____

年　月　日

第二部分

毕业论文开题报告

浙 江 大 學

本 科 生 毕 业 论 文

文献综述和开题报告

学生姓名 张乔

学生学号 3200102817

指导教师 纪守领

年级与专业 2020级 计算机科学与技术

所在学院 计算机科学与技术学院

一、题目：

二、指导教师对文献综述、开题报告、外文翻译的具体要求：

指导教师（签名）_____

年　月　日

一、文献综述

1　背景介绍

正文格式与具体要求**zjuthesisrules**

1.1　小节

1.1.1 小节

2　国内外研究现状

2.1　研究方向及进展

2.2　存在问题

3　研究展望

二、开题报告

1　问题提出的背景

正文格式与具体要求zjuthesisrules

1.1　背景介绍

1.1.1 项目提出的原因

1.2　本研究的意义和目的

2　项目的主要内容和技术路线

2.1　主要研究内容

2.2　技术路线

2.3　可行性分析

3　研究计划进度安排及预期目标

3.1　进度安排

3.2　预期目标

三、外文翻译

摘要

动机：药物反应预测（DRP）在精准医疗中扮演着重要角色（例如用于癌症分析和治疗）。深度学习算法的最新进展使得基于基因档案准确预测药物反应成为可能。然而，现有方法忽略了基因之间潜在的关系。此外，细胞系/药物之间的相似性很少被明确考虑。

结果：我们提出了一个新颖的药物反应预测框架，称为 TGSA，以更好地利用先前的领域知识。TGSA 由用于药物反应预测的双图神经网络（TGDRP）和一个相似性增强（SA）模块组成，用于融合细粒度和粗粒度信息。具体来说，TGDRP 基于 STRING 蛋白质-蛋白质关联网络将细胞系抽象为图，并使用图神经网络（GNNs）进行表示学习。SA 将 DRP 视为异构图上的边回归问题，并利用 GNNs 平滑相似细胞系/药物的表示。此外，我们引入了一个辅助的预训练策略，以弥补数据稀缺和分布外泛化能力差的已识别限制。在 GDSC2 数据集上的广泛实验表明，我们的 TGSA 在各种实验设置下始终优于所有最先进的基线。我们通过消融实验进一步评估了 TGSA 的每个组成部分的有效性和贡献。TGSA 的优异表现显示了其在精准医疗临床应用中的巨大潜力。

1 引言

精准医疗旨在根据个体基因、生活方式和环境为每位患者定制治疗方案[1]。尽管在癌症分析中，基因组学往往与精准医疗同义，但基于遗传信息实现精准医疗仍然面临挑战，因为基因型与表型之间的确切关系尚未明确（Friedman et al., 2015）。这些挑战促使研究人员进行大规模的抗癌药物筛选，并提出计算方法来探索患者遗传档案对药物反应的潜在影响（Baptista et al., 2021）。

最近，高通量筛选技术的发展促进了对一系列癌细胞系进行数百种抗癌药物的筛选。已经启动并系统化为公共库的几个大规模癌症遗传项目，包括癌症细胞系百科全书（CCLE）（Barretina et al., 2012）和癌症药物敏感性遗传学（GDSC）（Yang et al., 2013）。鉴于这些数据集不仅提供了丰富的抗癌药物反应信

息，还提供了癌细胞系的遗传档案，它们在促进药物反应预测（DRP）技术的发展方面大有裨益。

在过去的十年中，提出了两类主要的计算 DRP 方法。(i) 矩阵分解：这些方法通常将药物反应矩阵分解为两个低秩矩阵，表示细胞系和药物的表示。通过扩展矩阵分解，学习通路-药物反应关联（Ammad-ud din et al., 2016）和细胞系/药物相似性（Suphavilai et al., 2018; Wang et al., 2017）以提高性能。虽然这些早期尝试在 DRP 方面取得了显著进展，但它们没有探索药物和细胞系的特征，并且在推断未见药物或细胞系的药物反应时表现出较差的外推性。(ii) 机器学习：许多研究尝试用机器学习算法解决上述问题，例如弹性网络（Barretina et al., 2012）、支持向量机（Dong et al., 2015）和随机森林（Iorio et al., 2016）。虽然这些方法显示出相当的效果，但它们仍然存在主要缺点。首先，常用的特征，如药物的分子指纹和细胞系的遗传档案，是高维数据，可能导致方法严重过拟合（Teschendorff, 2019）。其次，特征选择是不可或缺的，但需要深入了解涉及的生物过程（Mayr et al., 2016）。第三，生物数据中的关系过于复杂，无法完全由这些容量有限的方法建模。

凭借在表示高维数据（如图像、视频和文本）方面的前所未有的能力，深度学习方法吸引了众多关注。最近，深度学习也找到了进入 DRP 的道路，并超越了传统的机器学习算法（Baptista et al., 2021）。根据训练模式，基于深度学习的 DRP 方法可以分为（i）两阶段框架和（ii）端到端框架。两阶段框架通常首先通过自编码器学习压缩表示，然后基于压缩表示进行 DRP（Chiu et al., 2019; Ding et al., 2018）。端到端框架从原始数据学习高级表示，并同时推断药物反应。具体来说，MOLI（Sharifi-Noghabi et al., 2019）使用三种特定类型的子网络从多组学数据中学习表示，并将学习到的表示连接作为输入到后续子网络进行 DRP。CDRscan（Chang et al., 2018）使用卷积神经网络（CNNs）分别从分子指纹和体细胞突变中学习药物和细胞系的表示。此外，CNNs 已被用于处理药物的 SMILES（Liu et al., 2019）和 Kekule 结构（Corte´s-Ciriano and Bender, 2019）。

利用药物的 2D 拓扑结构，药物可以明确表示为分子图，其中节点和边分别表示原子和化学键。随着深度学习方法在图问题上的发展，GNNs 已被应用于分

子表示学习，并成为药物发现的最先进方法（Sun et al., 2020）。最近，DRP 的尝试也密切关注了使用 GNNs 处理药物。DeepCDR（Liu et al., 2020）提出了一种统一的图卷积网络（UGCN），以捕捉药物的内在化学结构。然而，很少有研究对应用 GNNs 学习细胞系的表示进行研究。

我们总结了现有基于深度学习的 DRP 方法仍然面临的主要局限性：

- 现有方法尚未能够捕捉基因之间的关系，这对于准确表示细胞系至关重要。例如，CNNs 由于其探索图像空间拓扑的能力而在计算机视觉中已成为领先的架构（Lecun et al., 1998）；但是，CNNs 仍然不适合处理没有易于利用的空间信息的基因组档案。

- 现有方法没有充分利用细胞系/药物之间的相似性。请注意，具有相似遗传档案的细胞系和具有相似化学结构的药物通常具有相似的药物反应，如先前研究讨论的那样（Wang et al., 2017; Zhang et al., 2015）。

- 现有方法在盲测试中对未见细胞系和药物的泛化仍然面临困难，这意味着这些方法远远不足以满足临床癌症应用中的精准医疗。

为了解决上述局限性，我们提出了一个新颖的药物反应预测框架，称为具有相似性增强的双图神经网络（TGSA）（见图 1 概述）。TGSA 由用于药物反应预测的双图神经网络（TGDRP）和一个相似性增强（SA）模块组成，结合先前的领域知识来融合细粒度（基因级别和原子级别）和粗粒度（样本级别）信息。TGDRP 利用两个 GNN 编码器分别基于分子图和 STRING 蛋白质-蛋白质关联网络（Szklarczyk et al., 2019）学习药物和细胞系的表示，然后将学习到的表示连接并输入到全连接网络（FCN）进行最终预测。在 SA 模块中，我们通过将 DRP 视为异构图上的边回归问题来融合细胞系/药物之间的相似性，其中每个节点表示药物或细胞系。SA 利用 GNNs 平滑相似细胞系/药物的表示。此外，我们为 TGDRP 的药物编码器引入了一个辅助预训练策略，以缓解数据不足和分布外分子泛化能力差的问题。我们的主要贡献如下。

- 更好地利用先前的领域知识。我们提出了 TGDRP，将基因之间的关系纳入

考虑，这对于捕捉细胞系的复杂模式至关重要。我们还提出了 SA，将药物和细胞系之间的相似性纳入考虑。据我们所知，我们是第一个在 DRP 中使用 GNNs 学习细胞系表示的。

- 针对有限数据的策略。我们适应了一个辅助预训练策略，以进一步提高泛化能力，特别是在分布外药物上的泛化能力。
- 在不同粒度的信息融合。TGDRP 和 SA 被耦合到 TGSA 中，以融合细粒度和粗粒度信息。
- 有希望的性能。我们的综合实验表明，在各种实验设置下，TGSA 在 GDSC2 数据集上的表现优于所有基线，实证显示了 TGSA 的卓越预测能力。

2 材料和方法

2.1 问题定义

在这项工作中，我们将 DRP 定义为对应药物-细胞系对的对数归一化半最大抑制浓度（$\ln(\text{IC}_{50})$）值的回归问题，定义映射函数 $f : D \times C \to Y$，其中 $D = \{d_1, d_2, \ldots, d_n\}$ 和 $C = \{c_1, c_2, \ldots, c_m\}$ 分别代表药物和细胞系的集合，$Y \in \mathbb{R}^{m \times n}$ 是药物反应矩阵。矩阵中的每个元素 Y_{ij} 表示细胞系 c_i 与药物 d_j 的 $\ln(\text{IC}_{50})$ 值。

2.2 数据准备

在本节中，我们描述了如何将三个公共数据库：GDSC2、CCLE 和 COSMIC（Tate et al., 2019）整合到我们的实验数据集中。此外，我们展示了如何将药物和细胞系表示为图形结构。

数据整合 GDSC2 提供了数百个细胞系和药物的 $\ln(\text{IC}_{50})$ 值。CCLE 为细胞系提供了详尽的遗传档案，包括基因表达（EXP）、体细胞突变（MU）和拷贝数变异（CNV）。此外，我们依据 COSMIC 数据库（Tate et al., 2019）选择了 706 个

与癌症相关的基因进行特征选择。特别的，没有 PubChem ID 的药物以及缺乏上述任何一类组学数据的细胞系被排除。因此，我们得到了 580 个细胞系和 170 种药物，涉及 82833 个有效的 $\ln(IC_{50})$ 值（参见补充材料表 S1-S3），其中大约 16%（15767 个）是未知的。GDSC2 数据的基本统计信息在补充材料图 S1 中展示。

图构建方法 一般来说，一个图 $G = (V, E)$ 可以用元组 (A, F) 表达，其中集合 V 代表 N 个节点，集合 E 代表边，矩阵 $A \in \{0,1\}^{N\times N}$ 表示邻接矩阵，而 $F \in \mathbb{R}^{N\times K}$ 代表节点特征向量的矩阵。在我们的工作中，每个分子图 G_d 是根据 Liu et al. (2020) 的方法构建的，所有原子的特征细节都展示在表 1 中。对于每个细胞系图 $G_c = (V_c, E_c)$，V_c 和 E_c 分别代表基因及其相互作用，$\mathbf{F}_c \in \mathbb{R}^{706\times 3}$ 由上述基因水平的多组学数据决定。不同于分子中的化学键，细胞系中的基因间并无明显的结构关系，因此，我们引入了基因相互作用网络作为先验知识以定义 $\mathbf{A}_c \in \mathbb{R}^{706\times 706}$。STRING 数据库提供了经过策划的基因相互作用信息（Szklarczyk et al., 2019）。具体来说，若基因 i 与基因 j 在 STRING 中的联合相互作用得分超过预设的阈值 s（默认值设为 0.95），则设置 $\mathbf{A}_{c_{ij}} = 1$，否则为 0。从统计角度看，G_c 是一张断点图，最大的连通分量包含 408 个节点，其稀疏度为 0.006，平均度为 4.46。

2.3 我们提出的方法

不失一般性，我们首先简要回顾消息传递神经网络（MPNNs）[2]作为一个典型的框架，以介绍图神经网络（GNNs）的基本概念。接下来，我们介绍 TGDRP 和 SA（图 1）。最后，我们讨论如何将 TGDRP 和 SA 结合到 TGSA 中。

MPNNs MPNNs 的主要思想是，每个节点可以在消息传递阶段递归地从其邻居节点接收消息。在读出阶段，MPNNs 使用一个排列不变的读出函数将所有节点表示汇总成固定长度的图级表示，这些表示可用于各种下游任务。正式地说，给定一个图 G，其节点表示 $h_v \in \mathbb{R}^n$ 对于节点 v 和图级表示 z_l 在层 l 可以如下计算：

$$m_v^{(l)} = \sum_{w \in \mathcal{N}(v)} M^{(l)}\left(h_v^{(l-1)}, h_w^{(l-1)}, e_{vw}\right), \quad (1) \tag{2-1}$$

$$h_v^{(l)} = U^{(l)}\left(h_v^{(l-1)}, m_v^{(l)}\right), \quad (2) \tag{2-2}$$

$$z_l = R\left(\{h_v^{(l)} \mid v \in G\}\right), \quad (3) \tag{2-3}$$

其中 $m_v^{(l)}$ 表示节点 v 的邻居在层 $l-1$ 的总消息，$\mathcal{N}(v)$ 表示节点 v 的邻居集合，e_{vw} 表示节点 v 和 w 之间的边的特征，M、U 和 R 分别是消息函数、更新函数和读出函数。

TGDRP 为了探索和利用细胞系图的拓扑结构，更好地学习细胞系表示，我们提出了基于 GNN 的 TGDRP。如图 1 所示，TGDRP 是一个双分支网络，它接受一个分子图 G_d 和一个细胞系图 G_c 作为输入，并输出预测的 $\ln(\text{IC}_{50})$。TGDRP 包括以下三个组成部分。

药物分支 给定一个分子图 G_d，我们使用图同构网络（GIN）[Xu2019] 作为 GNN_d 来更新原子特征。GIN 在一系列与分子相关的任务上取得了最先进的性能[Hu2020]。受到跳跃知识网络的启发，以在不同尺度上保留信息[3]，我们通过全局最大池化读出的层间图级表示进行连接，随后是 FCN_d 来学习药物表示 $z_d \in \mathbb{R}^{256}$。

细胞系分支 给定一个细胞系图 G_c，我们使用图注意力网络（GAT）[Velickovic2018] 作为 GNN_c 来更新基因特征。在消息传递阶段，GAT 通过自注意力机制为不同的邻居节点分配不同的权重。细胞系图包含关于基因相互作用的层次信息[4]。但是，普通的 GAT 和全局池化本质上是平坦的，无法捕捉这种信息[5]。因此，我们使用 Graclus[6] 在每层 GAT 之后逐渐将图粗化，通过聚类两个节点成一个“超级节点”。基于谱聚类和核 k 均值，Graclus 是一个高效且无模型的图聚类算法[Bianchi2020]。先前的研究证明了 Graclus 在分组强连接基因方面的有效性。通过这种操作，不仅捕获了细胞系图的层次结构，还保留了 CNN 中池化操作的优势。细胞系图有两个有益的特征：(i) 所有细胞系图共享相同的拓扑结构。在实践中，我们只需要在训练前运行一次 Graclus，以避免重复计算。(ii) 与分子图相比，细

胞系图的规模更大。我们直接连接所有超级节点表示来读出图级表示，而不是全局池化，后者只保留了一阶统计信息并丢失了许多有用信息。通过图级表示，FCN_c 随后输出细胞系表示 $z_c \in \mathbb{R}^{256}$。

预测 FCN 药物和细胞系表示被连接起来，并输入预测 FCN，产生最终预测的 $\ln(\text{IC}_{50})$。

预训练策略 预训练策略在计算机视觉[7] 和自然语言处理**Kenton2019** 中已经被广泛研究，但它们在 DRP 中的效果尚未得到很好的探索。由于以下两个问题，训练有效的 GNNs 来捕捉药物的特征是一个巨大的挑战：(i) 数据有限：整个数据集中只有 170 种药物。(ii) 分布外样本：测试集中的药物图在结构上与训练集中的药物图非常不同，导致分布外泛化能力差。为了缓解这些问题，我们在大规模分子数据集中预训练 GNN_d，并将化学领域的知识转移到我们的任务中。在这项研究中，我们遵循中的预训练策略，其关键思想是在节点和图级别捕获领域特定的语义。具体来说，我们首先在 ZINC15 数据集[8] 上进行节点级别的自监督预训练，使用深度图信息最大化（DGI）**Velickovic2019**。然后，我们在包含 456k 分子和 1310 种生化特性的 ChEMBL 数据集[9] 上进行图级多任务监督预训练。在预训练期间，GNN_d 积累了分子的化学知识（例如价性和生化特性）。与不同，我们随机保留一个验证集进行模型选择。

相似性增强 正如在第 1 节中讨论的，基于深度学习的药物反应预测（DRP）方法存在已知限制，那就是相似的细胞系或药物倾向于呈现相似的表示，进而引导相似的反应。在矩阵分解方法中，细胞系或药物的相似性通常被用作正则化项，目的是为了提升 DRP 性能[10-11]。然而，这种方式的外推性能较差，并且它只尝试减少相细胞系或药物的表示之间差异，却忽略了它们之间更深层次的相互关系。作为低通滤波器的图神经网络（GNNs）有助于使得图中相邻节点呈现出相似的表示[12]，并在节点间传递信息[2]。因此，我们在算法设计中考虑到了细胞系或药物间的相似性，并利用 GNN 来解决上述问题。我们具体地将 DRP 框架当作一个异构图上的边回归问题，并运用 GNN 来平滑化细胞系或药物的节点表示。

这便形成了我们特有的相似性增强（SA）模块。该异构图是根据已知的药物与细胞系之间的反应连接信息来构建的，每一种药物或细胞系都会与另外若干个最相似的实例相连接，形成了一个同质的 k-最近邻图。具体而言，这样一个异构图可以表示为 $G = (V, E)$，其中 $V = D \cup C$，且 $E = E_{IC} \cup E_d \cup E_c$。$E_d$ 和 E_c 是基于扩展连接指纹（Extended Connectivity Fingerprints, ECFP）的 Jaccard 相似性度量以及基因表达数据的 Pearson 相关系数计算得到的[13]。在 E_{IC} 中，则表示那些具有相应的 $\ln(IC50)$ 标签的药物与细胞系之间的连接。对于每一种细胞系或药物，我们将类型特定的 GraphSAGE 方法采用在异构图 G 上，利用邻居的信息来更新它们的表示。这样的处理使得我们的节点表示富含信息，且那些相似的细胞系或药物天然地倾向于获取相似的节点表示。SA 模块的计算过程可以如下描述：

$$Z_c = \text{GraphSAGE}_c(X_c, E_c), \quad Z_d = \text{GraphSAGE}_d(X_d, E_d)$$

这里的 $\text{GraphSAGE}_c(\cdot)$ 与 $\text{GraphSAGE}_d(\cdot)$ 指的是分别作用于细胞系和药物的特化 GraphSAGE 模型（详细公式可参见补充材料）。$X_c \in \mathbb{R}^{m \times \dim_c}$ 和 $X_d \in \mathbb{R}^{n \times \dim_d}$ 分别表示细胞系和药物的原始特征矩阵，且每一行对应一个独立的细胞系或药物实例。$Z_c \in \mathbb{R}^{m \times 256}$ 和 $Z_d \in \mathbb{R}^{n \times 256}$ 则分别是作为输出的细胞系和药物的表示矩阵。细胞系的输出表示为 Z_c^i，而药物的则为 Z_d。这一预测步骤在 TGDRP 中与我们的 SA 模块是相同的。

TGSA 由于 TGDRP 中的节点和 SA 中的节点表示不同粒度级别上的对象（即基因与细胞系，原子与药物），TGDRP 捕获基因级别和原子级别的信息，而 SA 捕获样本级别的信息。为了融合这样的细粒度和粗粒度信息，我们寻求通过微调来耦合 TGDRP 和 SA。首先，我们端到端地训练 TGDRP。然后在 SA 中，X_d 和 X_c 通过将相应的训练好的 GNN_d 和 GNN_c 应用于分子图和细胞系图集合来生成。为了避免额外的参数，类型特定的 GraphSAGE 和预测 FCN 的隐藏维度和隐藏层数与 TGDRP 中的相应 FCN 保持相同。最后，对于微调，GraphSAGE 和预测 FCN 的参数由 TGDRP 中的相应训练 FCN 初始化。通过这些过程，TGDRP

可以被视为 TGSA 的一个特殊情况，其中 $E_d \cup E_c = \phi$。

2.4 实验设置

为了全面展示我们方法的有效性，我们旨在检查以下三个问题：

- Q1: TGSA 在不同的实验设置下能否超越最先进的基线？
- Q2: TGSA 的每个组成部分是否有效？
- Q3: TGSA 能否捕获任何生物学意义？

根据[14] 的建议，我们在两种实验设置下评估我们的方法。第一种是重新发现已知的药物-细胞系反应：整个数据集通过分层抽样分为训练/验证/测试集，比例为 8:1:1，每个实验使用 10 个随机种子重复。第二种是盲测试（留一药物/细胞系）：整个数据集在细胞系/药物级别上分割，以确保测试集仅包括训练阶段中未见的药物/细胞系，并且进行 5 折交叉验证，其中三折被视为训练集，另外两折被视为验证集和测试集。这种更严格的场景符合药物重定位和推荐的临床应用。考虑到测试集的数据分布可能与训练集非常不同，交叉验证可以更准确地估计真正的泛化误差。

我们比较了几种代表性的最先进方法，包括 CDRscan[15]，MOLI[SharifiNoghabi2019]，和 tCNNS[16]，以及两个最近的基于 GNN 的方法：GraphDRP[Nguyen2021] 和 DeepCDR[17]。为了探索细胞系图和 GNN_c 的有效性，我们对 DeepCDR 进行了一些小的调整以进行公平比较。具体来说，DeepCDR 的 UGCN 被我们的药品分支替换，DeepCDR 的多组学数据也被我们的替换。我们将这个变体称为 DeepCDR'。我们采用回归任务中广泛使用的三个指标来衡量性能：均方根误差（RMSE），平均绝对误差（MAE）和皮尔逊相关系数（r）。为了训练我们的模型，我们使用均方误差作为损失函数。所有基线和我们的模型都是在 PyTorch[18] 和 PyTorch Geometric[FeyandLenssen2019] 中

3 实验结果

3.1 与基线的比较

表 2 比较了我们的方法与最先进基线的架构和性能。我们的方法在所有三个指标上显著优于所有基线，从而对 Q1 提供了积极的答案。我们在不同的实验设置下讨论结果如下。

重新发现已知的药物-细胞系反应 tCNNS 和 GraphDRP 达到了可比的性能，但是 GraphDRP 的稳健性较差，标准差较高。这一现象与 GNNs 在许多与分子相关的任务上优于传统方法的通常观察不一致[19]。我们推测这可能是因为有限的药物数量限制了 GNNs 学习到富有表现力和稳健的表示的能力。DeepCDR' 在性能上显著优于其他基线。我们认为，DRP 从多组学数据中受益，这是 DeepCDR' 和 GraphDRP 之间的主要区别。因此，我们建议在可能的情况下，使用相关的组学数据来表示细胞系，因为这些数据可以在不显著增加计算成本的情况下稳定地提高性能。

与 DeepCDR' 相比，我们的方法 $\text{TGDRP}_{\text{w/o pre}}$ 在 RMSE 上降低了 0.037，在 MAE 上降低了 0.028，相对改进大约为 4.1%。图 2 展示了 TGSA 的预测结果。具体而言，图 2c 展示了观测到的 $\ln(IC50)$ 与预测的 $\ln(IC50)$ 之间的关系，这两者显示出高度的线性相关性。图 2a 和 2b 可视化了在细胞系/药物级别上测试 RMSE 的分布，可以看到这些分布近似于正态分布。与细胞系相比，有几种药物在泛化性能上表现特别差，其 RMSE 超过了 3。这一现象可能是因为在巨大的化学空间中，测试药物超出了分布。

盲测试 如图 2e 和 2f 所示，在盲测试场景中，尽管所有方法的表现都不如重新发现已知的药物-细胞系反应场景，特别是在留一药物外的情况下，但我们的方法 $\text{TGDRP}_{\text{w/o pre}}$ 仍然表现优于最佳基线 DeepCDR'。需要注意的是，我们没有比较所有变体的性能，因为预训练和 SA 模块只分别明确影响药物和细胞系的表示（详见第 3.2 节）。在留一药物外的场景中，$\text{TGDRP}_{\text{w/o pre}}$ 在皮尔逊相关系数（r）

方面比 DeepCDR' 高出 0.033 (0.493 对比 0.46)，而整个 TGDRP 达到了较高的 r 值，为 0.527。性能下降表明，在泛化到结构不同的药物方面，我们的方法仍然面临困难。预训练策略通过利用大规模分子结构数据来缓解这一缺陷；然而，所见改进仍不够显著。我们认为，这可能是因为抗癌药物是一类非常特殊的分子，通用领域知识似乎不足以应对这一挑战。在留一细胞系外的场景中，TGDRP 和 TGSA 略显优于 DeepCDR'，Pearson 相关系数（r）分别为 0.874 和 0.872。我们的实验结果显示，与留一药物相比，TGSA 在留一细胞系外的场景中显示出潜力，表明 TGSA 的预测可以作为药物推荐的一个可靠参考。

3.2 消融实验

为了回答 Q2，我们从三个角度进行了消融实验：基因特征、细胞系图的拓扑结构和框架组件。为了进行公平比较，我们也对 DeepCDR' 进行预训练（记为 DeepCDR' pre）。

基因特征 评估三种类型的组学数据，即基因表达（EXP）、体细胞突变（MU）和拷贝数变异（CNV）的贡献，是非常必要的。我们在细胞系图中只使用一种类型的组学数据并排除其他类型（$F_c \in \mathbb{R}^{706\times 1}$）来进行评估。如表 3 上半部分所示，使用单一组学数据的 TGDRP，特别是 $\text{TGDRP}_{\text{CNV}}$，在所有单组学变体中展现了最佳性能，甚至超越了带有预训练的 DeepCDR'（DeepCDR' pre）。这强调了 TGDRP 在处理单一组学数据方面的强大能力。

拓扑结构 为了评估基于 STRING 的细胞系图拓扑结构的影响，我们比较了 TGDRP 模型在具有不同拓扑结构时的性能。具体来说，我们考察了以下几种拓扑：由 Erdős-Rényi 模型生成的随机图[Erdos1960]、根据细胞系特征的排列图、以及用基因表达的 Pearson 相关系数替换 STRING 交互得分得到的统计图。注意，为了进行公平比较，这些拓扑结构的平均度数被设置为与基于 STRING 的拓扑结构相同。结果表明，搭配 Erdős-Rényi 随机图的 TGDRP（即无意义拓扑的 TGDRP）与 DeepCDR' pre 的表现相似，但是使用 STRING 拓扑的模型在所有对比组中均

具有最佳表现，这突显了 STRING 数据中潜在有益知识的价值。

框架组件 预训练策略和 SA 在以前的 DRP 方法中很少被考虑。为了验证它们在我们方法中的效果，我们进行了消融实验，比较了两个 TGSA 变体：$\text{TGDRP}_{\text{w/o pre}}$（即没有使用预训练策略的 TGDRP）和 TGDRP。它们的性能如表 2 下半部分所示。考虑到我们的 SA 模块，我们发现只有通过细胞系相似性增强才能实现均方根误差（RMSE）的最大降低（0.008），这表明药物相似性增强似乎无用。这一现象与最近的研究[11]一致，而与 Zhang et al. 的研究结果相反[20]。我们怀疑药物反应相似性过于复杂，无法通过 ECFP 的 Jaccard 相似性来衡量。图 2g 和 h 展示了带有和不带 SA 的来自前 5 种最常见组织类型的细胞系的 t-SNE[vanderMaaten2008] 2D 表示，这证实了 SA 的效果，因为 SA 使得来自不同组织的细胞系更容易区分。同样，按 TCGA 癌症类型着色的可视化结果展示在补充图 S2 中。关于我们的预训练策略，TGDRP 的 RMSE 降低了 0.009。我们指出一个意外发现，即仅考虑节点级预训练比结合图级预训练可以获得更好的性能。我们推测，这种负面迁移可能是由于药物反应和 ChEMBL 中众多且模糊的生化检测之间的不一致性所导致的。

3.3 生物学意义

一个成功的 DRP 方法不仅应该做出准确的预测，还应该捕捉到重要的生物学意义。因此，我们进一步利用事后探索性策略和基于文献的案例研究来展示我们方法在发现未知药物-细胞系反应和寻找药物-基因相互作用方面的能力。发现未知药物-细胞系反应。在所有缺失药物-细胞系对的预测 $\ln(\text{IC}_{50})$ 中（见补充表 S5），Daporinad, EU-3 对具有最低的值（-8.05），其中 EU-3 是一种 B 细胞前体白血病细胞系，表明 Daporinad 对血液恶性肿瘤细胞具有优异的生物活性。这一预测得到了[21]的支持，他们报告称，由于 Daporinad 处理导致的 NAD 耗竭，大多数血液癌细胞对低浓度的 Daporinad 敏感，从而引发肿瘤细胞死亡。图 2d 展示了按药物分组的缺失药物-细胞系对的预测 $\ln(\text{IC}_{50})$ 的分布，其中 Bortezomib 具有最低的中位数。这一观察与 Bortezomib 对多种癌症（如复发性多发性骨髓

瘤[22]和乳腺癌PeriyasamyThandavan2010）具有治疗效果的实验结果一致。药物-基因相互作用。给定一个药物-细胞系对，我们修改了 Grad-CAM[23]来估计节点级基因重要性，因为我们的方法中基因被表示为节点。具体来说，将 ReLU 函数替换为绝对函数，以适应我们的回归问题，我们不仅关注对 $\ln(\mathrm{IC}_{50})$ 有正影响的特征，正如原始的 Grad-CAM 是为分类问题提出的。因此，Grad-CAM 的公式可以重写为：

$$L_{l,n} = \sum_{k=1}^{K} \left| \frac{\partial y}{\partial h_{k,n}} \right| h_{k,n}$$

其中 $L_{l,n}$ 表示相应图中节点 n 的重要性得分。

根据计算出的基因重要性得分，几个案例的前 5 个重要基因如表 4 所示。值得注意的是，我们发现许多排名靠前的基因已经被证实与相应药物的作用机制有关。例如，PAFAH1B2 是 HIF-1α 的靶基因[24]，后者已被报道通过抑制 Bortezomib 抑制肿瘤对低氧的适应[25]。对于（Bortezomib, 8MGBA）对，PAFAH1B2 排名第一。此外，Irinotecan 是 Camptothecin 的衍生物，针对 DNA 复制途径[26]。在它们对 KMH2 细胞系的敏感性中，靶基因 TOP1 和与 DNA 复制相关的基因 POLQ（DNA 聚合酶 η）分别排名第四和第三。此外，BRD4 已被报道是 AZD5153 的潜在靶点，通过调节转录程序[27]，并且我们的方法估计它在 AZD5153 对 NCIH526 细胞系的敏感性中很重要。

总之，这些基于文献的案例研究表明，我们的 TGSA 可以应用于发现新的治疗效果以及抗癌药物的潜在治疗靶点。

4 讨论

在本文中，我们提出了一个新颖的药物反应预测框架，称为 TGSA，它结合了先前的领域知识来融合细粒度和粗粒度信息。为了探索基因之间的关系，我们结合了基于 STRING 的蛋白质-蛋白质关联网络来构建细胞系图。为了捕捉细胞系/药物之间的相似性，我们将 DRP 视为异构图上的边回归问题，并利用 GNNs 平滑相似细胞系/药物的表示。此外，我们引入了一个辅助预训练策略，以缓解数据有限的困境，并提高药物的分布外泛化能力。在 GDSC2 数据集上的广泛实

验表明，我们的 TGSA 在不同的实验设置下优于几种最先进的 DRP 方法。消融研究验证了 TGSA 的每个组成部分的有效性和贡献。此外，生物学案例分析表明，TGSA 可以在一定程度上启发专家进一步探索相关领域知识。尽管上述表现充满希望，我们的方法仍需解决几个局限性，并为未来研究提供了见解。首先，细胞系图的拓扑结构仅基于 STRING 的联合相互作用得分简单地构建。为了专门确定图拓扑，考虑更丰富的领域知识是一个富有成效的途径。其次，Graclus 只考虑图拓扑，而忽略了节点特征[Bianchi2020]，因此细胞系图不能根据特定的 DRP 任务自适应地粗化。值得考虑的是应用基于特征的层次池化算法来帮助 GNNs 学习更复杂的细胞系表示。最后，盲测试场景的性能仍然不尽人意，尤其是在留一药物外的场景中。需要进一步努力来提高 GNNs 的分布外泛化能力，研究药物构象感知的自监督学习对这方面看起来很有前景。总之，我们的 TGSA 将化学领域知识整合到深度学习中，TGSA 的有希望的性能显示出其在精准医疗的临床应用中的巨大潜力。

5 外文翻译参考文献

[1] HODSON R. Precision Medicine[J]. NATURE, 2016, 537(7619): S49. DOI: 10.1038/537S49a.

[2] GILMER J, SCHOENHOLZ S S, RILEY P F, et al. Neural Message Passing for Quantum Chemistry[C]//PRECUP D, TEH . Proceedings of Machine Learning Research: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, VOL 70: vol. 70. 2017.

[3] XU K, HUANG B, LIU X, et al. A Low-Power Pyramid Motion Estimation Engine for 4K@30fps Realtime HEVC Video Encoding[C]//IEEE International Symposium on Circuits and Systems: 2018 IEEE INTERNATIONAL SYMPOSIUM ON CIRCUITS AND SYSTEMS (ISCAS). IEEE; Politecnico Torino; Analog Devices; CADENCE; Texas Instruments; AiCTX; Springer Nat; River Publishers; Nat Electron; CAS, 2018. DOI: 10.1109/ISCAS.2018.8350934.

[4] ERWIN D H, DAVIDSON E H. The Evolution of Hierarchical Gene Regulatory Networks [J]. NATURE REVIEWS GENETICS, 2009, 10(2): 141-148. DOI: 10.1038/nrg2499.

[5] YING R, YOU J, MORRIS C, et al. Hierarchical Graph Representation Learning with Differentiable Pooling[C]//BENGIO S, WALLACH H, LAROCHELLE H, et al. Advances in Neural Information Processing Systems: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 31 (NIPS 2018): vol. 31. 2018.

[6] DHILLON I S, GUAN Y, KULIS B. Weighted Graph Cuts without Eigenvectors: A Multilevel Approach[J]. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 2007, 29(11): 1944-1957. DOI: 10.1109/TPAMI.2007.1115.

[7] HE K, GIRSHICK R, DOLLAR P. Rethinking ImageNet Pre-Training[C]//IEEE International Conference on Computer Vision: 2019 IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV 2019). IEEE; IEEE Comp Soc; CVF, 2019: 4917-

4926. DOI: 10.1109/ICCV.2019.00502.
[8] STERLING T, IRWIN J J. ZINC 15-Ligand Discovery for Everyone[J]. JOURNAL OF CHEMICAL INFORMATION AND MODELING, 2015, 55(11): 2324-2337. DOI: 10.1021/acs.jcim.5b00559.
[9] GAULTON A, BELLIS L J, BENTO A P, et al. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery[J]. NUCLEIC ACIDS RESEARCH, 2012, 40(D1): D1100-D1107. DOI: 10.1093/nar/gkr777.
[10] GUAN N N, ZHAO Y, WANG C C, et al. Anticancer Drug Response Prediction in Cell Lines Using Weighted Graph Regularized Matrix Factorization[J]. MOLECULAR THERAPY-NUCLEIC ACIDS, 2019, 17: 164-174. DOI: 10.1016/j.omtn.2019.05.017.
[11] WANG L, LI X, ZHANG L, et al. Improved Anticancer Drug Response Prediction in Cell Lines Using Matrix Factorization with Similarity Regularization[J]. BMC CANCER, 2017, 17(513). DOI: 10.1186/s12885-017-3500-5.
[12] WU F, ZHANG T, DE SOUZA A H, Jr., et al. Simplifying Graph Convolutional Networks[C]//CHAUDHURI K, SALAKHUTDINOV R. Proceedings of Machine Learning Research: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, VOL 97: vol. 97. 2019.
[13] ROGERS D, HAHN M. Extended-Connectivity Fingerprints[J]. JOURNAL OF CHEMICAL INFORMATION AND MODELING, 2010, 50(5): 742-754. DOI: 10.1021/ci100050t.
[14] BAPTISTA D, FERREIRA P G, ROCHA M. Deep Learning for Drug Response Prediction in Cancer[J]. BRIEFINGS IN BIOINFORMATICS, 2021, 22(1, SI): 360-379. DOI: 10.1093/bib/bbz171.
[15] CHANG Y, PARK H, YANG H J, et al. Cancer Drug Response Profile Scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature[J]. SCIENTIFIC REPORTS, 2018, 8(8857). DOI: 10.1038/s41598-018-27214-6.
[16] LIU P, LI H, LI S, et al. Improving Prediction of Phenotypic Drug Response on Cancer Cell Lines Using Deep Convolutional Network[J]. BMC BIOINFORMATICS, 2019, 20(408). DOI: 10.1186/s12859-019-2910-6.
[17] LIU Q, HU Z, JIANG R, et al. DeepCDR: A Hybrid Graph Convolutional Network for Predicting Cancer Drug Response[J]. Bioinformatics (Oxford, England), 2020, 36(2): I911-I918. DOI: 10.1093/bioinformatics/btaa822.
[18] PASZKE A, GROSS S, MASSA F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library[C]//WALLACH H, LAROCHELLE H, BEYGELZIMER A, et al. Advances in Neural Information Processing Systems: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 32 (NIPS 2019): vol. 32. 2019.
[19] SUN M, ZHAO S, GILVARY C, et al. Graph Convolutional Networks for Computational Drug Development and Discovery[J]. BRIEFINGS IN BIOINFORMATICS, 2020, 21(3): 919-935. DOI: 10.1093/bib/bbz042.
[20] ZHANG N, WANG H, FANG Y, et al. Predicting Anticancer Drug Responses Using a Dual-Layer Integrated Cell Line-Drug Network Model[J]. PLOS COMPUTATIONAL BIOLOGY, 2015, 11(e1004498). DOI: 10.1371/journal.pcbi.1004498.
[21] NAHIMANA A, ATTINGER A, AUBRY D, et al. The NAD Biosynthesis Inhibitor APO866 Has Potent Antitumor Activity against Hematologic Malignancies[J]. BLOOD, 2009, 113(14): 3276-3286. DOI: 10.1182/blood-2008-08-173369.
[22] RICHARDSON , BARLOGIE B, BERENSON J, et al. A Phase 2 Study of Bortezomib in Relapsed, Refractory Myeloma[J]. NEW ENGLAND JOURNAL OF MEDICINE, 2003, 348(26): 2609-2617. DOI: 10.1056/NEJMoa030288.
[23] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization[C]//IEEE International Conference on Computer Vision: 2017 IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV). IEEE; IEEE Comp Soc, 2017: 618-626. DOI: 10.1109/ICCV.2017.74.
[24] MA C, GUO Y, ZHANG Y, et al. PAFAH1B2 Is a HIF1a Target Gene and Promotes Metas-

tasis in Pancreatic Cancer[J]. BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS, 2018, 501(3): 654-660. DOI: 10.1016/j.bbrc.2018.05.039.

[25] SHIN D H, CHUN Y S, LEE D S, et al. Bortezomib Inhibits Tumor Adaptation to Hypoxia by Stimulating the FIH-Mediated Repression of Hypoxia-Inducible Factor-1[J]. BLOOD, 2008, 111(6): 3131-3136. DOI: 10.1182/blood-2007-11-120576.

[26] POMMIER Y. Topoisomerase I Inhibitors: Camptothecins and Beyond[J]. NATURE REVIEWS CANCER, 2006, 6(10): 789-802. DOI: 10.1038/nrc1977.

[27] RHYASEN G W, HATTERSLEY M M, YAO Y, et al. AZD5153: A Novel Bivalent BET Bromodomain Inhibitor Highly Active against Hematologic Malignancies[J]. MOLECULAR CANCER THERAPEUTICS, 2016, 15(11): 2563-2574. DOI: 10.1158/1535-7163.MCT-16-0141.

四、外文原文

浙江大学本科生毕业论文（设计）编写规则

为规范我校本科生毕业论文(设计)编写格式,根据《学位论文编写规则》(GB/T 7713.1 - 2006)，结合我校实际情况，制定本本科生毕业论文（设计）编写规则。

1 本科生毕业论文（设计）工作文档

本科生毕业论文（设计）工作文档分两大部分。

第一部分(论文(设计)材料)编排顺序依次是前置部分、主体部分、结尾部分、《浙江大学本科生毕业论文（设计）任务书》、《浙江大学本科生毕业论文（设计）考核表》。

第二部分(开题材料)编排顺序依次是文献综述和开题报告封面、指导教师对文献综述和开题报告具体要求、目录、文献综述、开题报告、外文翻译和外文原文、《浙江大学本科生文献综述和开题报告考核表》。

本科生毕业论文（设计）工作文档的纸质版，可作为院（系）教学资料存档保存，参照上述编排顺序，打印装订成册，其中封面、题名页、承诺书、致谢、摘要、《浙江大学本科生毕业论文（设计）任务书》、 《浙江大学本科生毕业论文（设计）考核表》、指导教师对文献综述和开题报告具体要求、《浙江大学本科生文献综述和开题报告考核表》应单面打印，其它部分内容应双面打印；主体部分各章之间应分页。论文检测报告、浙江大学本科生毕业论文（设计）专家评阅意见、浙江大学本科生毕业论文（设计）现场答辩记录表等文档可作为附件单面打印装订在最后面。

本科生毕业论文（设计）工作文档的电子版，其内容中应不包含指导性、评价性及成绩考核等内容，如:《浙江大学本科生毕业论文（设计）任务书》、《浙江大学本科生毕业论文（设计）考核表》、指导教师对文献综述和开题报告具体要求、《浙江大学本科生文献综述和开题报告考核表》、论文检测报告、浙江大学本科生毕业论文（设计）专家评阅意见、浙江大学本科生毕业论文（设计）现场答辩记录表等。

1.1 前置部分

(1) 封面
(2) 题名页（可根据需要）
(3) 承诺书
(4) 勘误页(可根据需要)
(5) 致谢
(6) 摘要页

(7) 序言或前言(可根据需要)

(8) 目次页

(9) 图和附表清单（可根据需要）

(10) 符号、标志、缩略词、首字母缩写、计量单位、术语 等的注释表（可根据需要）

1.2 主体部分

(1) 引言（绪论）

(2) 正文

(3) 结论

1.3 结尾部分

(1) 参考文献

(2) 附录(可根据需要)

(3) 分类索引、关键词索引(可根据需要)

(4) 作者简历

2 编写规范与要求

2.1 语种要求

论文撰写语种，对于国际学生，参照 2017 年教育部、外交部、公安部联发的第 42 号令《学校招收和培养国际学生管理办法》执行；非国际学生遵照国家相关法律法规执行。

2.2 前置部分

2.2.1 封面

作者学号：全日制学生需要填写学号。

论文题目：应准确概括整个论文的核心内容，简明扼要，一般不能超过 25 个汉字，英文题目翻译应简短准确，一般不应超过 150 个字母，必要时可以加副标题。

2.2.2 承诺书

见浙江大学本科生毕业论文（设计）承诺书。

2.2.3 致谢

致谢对象限于对课题工作、毕业论文（设计）完成等方面有较重要帮助的人员。

2.2.4 摘要

包括中文摘要和英文摘要两部分。摘要应具有独立性和自含性，即不阅读论文全文就能获得必要信息。摘要的内容应包含与论文等同量的主要信息，供读者确定有无必要阅读

全文，也可供二次文献采用。摘要应说明研究目的、方法、结果和结论等，重点是结果和结论。不宜使用图、表、化学结构式、非公知公用的符号和术语。中文摘要的字数一般为300-600字以内，英文摘要实词在300个左右。英文摘要应与中文摘要内容相对应。摘要最后另起一行，列出3-8个关键词。关键词应体现论文特色，具有语义性，在论文中有明确的出处，并应尽量采用《汉语主题词表》或各专业主题词表提供的规范词。

2.2.5 序言或前言

毕业论文（设计）的序言或前言，一般是作者对本篇论文基本特征的简介，如说明研究工作缘起、背景、主旨、目的、意义、编写体例，以及资助、支持、协作经过等。这些内容也可在正文引言中说明。

2.2.6 目次页

论文中内容标题的集合。

2.2.7 图和附表清单

论文中如图表较多，可以分别列出清单置于目次页之后。图的清单应有序号、图题和页码。表的清单应有序号、表题和页码。

2.2.8 符号、标志、缩略词、首字母缩写、计量单位、术语等的注释表。

2.3 主体部分

包括引言（绪论）、正文和结论。

2.3.1 一般要求

2.3.1.1 引言（绪论）

应包括论文的研究目的、流程和方法等。论文研究领域的历史回顾、文献回溯、理论分析等内容，应独立成章，用足够的文字叙述。

2.3.1.2 正文

主体部分由于涉及不同的学科，在选题、研究方法、结果表达方式等有很大的差异，不能作统一的规定。但是，必须实事求是、客观真切、准备完备、合乎逻辑、层次分明、简练可读。

图：图包括曲线图、构造图、示意图、框图、流程图、记录图、地图、照片等。图应具有“自明性”；图宜有图题，即名称，置于图的编号后。图的编号和图题置于图下方；照片图要求主题和主要显示部分的轮廓鲜明，便于制版。如用 放大缩小的复制品，必须清晰，反差适中。照片上应有表示目的物尺寸的标度。

表：表应具有“自明性”。表宜有表题，即表的名称，置于表的编号之后。表的编号和表题应置于表上方。表的编排，一般是内容和测试项目由左至右横读，数据依序竖读。

毕业论文（设计）文献综述和开题报告考核

导师对开题报告、外文翻译和文献综述的评语及成绩评定：

成绩比例	文献综述 占（10%）	开题报告 占（15%）	外文翻译 占（5%）
分值			

导师签名 _____

年　月　日

学院盲审专家对开题报告、外文翻译和文献综述的评语及成绩评定：

成绩比例	文献综述 占（10%）	开题报告 占（15%）	外文翻译 占（5%）
分值			

开题报告审核负责人（签名/签章）_____

年　月　日